genpact
Transformation Happens Here

POINT OF VIEW

Online safety: Who is watching out for you?

AI and machine learning help manage user-generated content risks, but humans are still critical

Tide pods. Pizzagate. Putin's alleged interference in US elections. With more than four billion people accessing the internet today, there's little doubt that fake news is a powerful detriment to democracy and enterprise.

Consider: Every day, Facebook adds 500,000 new users - 6 new profiles every second. Five hundred million people visit Twitter each month without logging in. Over 95 million photos are posted on Instagram every day, and 300 hours of new YouTube videos are uploaded every minute. (Source: Brandwatch.com[1])

1. https://www.brandwatch.com/blog/amazing-social-media-statistics-and-facts/#section-2)%20%20

Too much of this user-generated content (UGC) is spam or phishing. With so much material out there, internet users need confidence that they are retrieving truly valuable nuggets of information.

They have reason for concern, given recent UGC abuses. Some obvious examples include:

- Charges of election interference have raised alarm around the world.
- A Washington Post investigation revealed that as many as 70% of product reviews could be fake.
- Facebook COO Sheryl Sandberg testified before the US Senate that the social media giant, in an "arms race" against misinformation, had deleted 583 million fake accounts in the first quarter of 2018 alone.
- YouTube hired 10,000 moderators in late 2017 when advertisers fled the platform after realizing that their ads were appearing next to inappropriate videos.

The ethical implications are significant, not to mention the obvious financial and reputation risks. People's trust in an institution is keenly associated with the web information linked to it.

In this environment, the management and filtering of UGC to eliminate non-conforming material is no longer an option. It's a must. The practice, defined as scanning and scrubbing out UGC for text, video, or images that contravene the values of a firm or institution, has become a prudent organization's defensive front line and nerve center.

As a result, growth in the content moderation market has been exponential.

We estimate that 500,000+ jobs will be created over next 3 years, supporting performance of new products impacting society, and not just for content moderation but overall for data-labeling in areas that range from virtual assistants to maps and online media libraries.

The need for established expertise

Organizations seeking to enter the fray - to gain the upper hand on content disseminated on their platforms - have reasonable questions: How do they protect their users? How do they protect their brands? How can companies verify that material they publish is trustworthy? What kind of guidelines should they devise to ensure the safety, credibility, and dependability of user-provided content? And how can they address the monumental task of applying those rules across a huge array of ads, reviews, news postings, discussion boards, and so on?

The fact is that they can't do it on their own. Few firms have the necessary skills, resources, or technology. One option is to turn to third-party Trust and Safety partners that have proven capabilities in policy setting, enforcement, and governance.

Not all methods of content moderation are equal. Those combining the work of human investigators with well-executed artificial intelligence (AI) and machine learning (ML), tend to be the most vigorous. Today, algorithms can identify published content that violates pre-determined set of rules or policies. But devising truly effective algorithms calls for a high level of expertise in AI and ML, given the immensity of the possible rules that need to be addressed and the constantly evolving nature of both UGC and objectionable content.

Tagging and flagging

Content moderation is a form of established tagging and flagging methods that advanced analytics experts have developed for other applications. Systems using this approach assess information to identify messages that meet certain criteria, then refer them to specialists who can verify actionable characteristics. Marketing specialists already use these techniques to identify content that mentions their products and services. Social scientists use them to track emerging behaviors and trends. Politicians use them to monitor voter concerns.

As the field evolves, five types of content moderation have emerged.

Types of content moderation

- **Pre-moderation** - the practice of clearing for approval UGC material before posting it. While this reduces institutional risk, the delay also reduces user satisfaction

- **Post-moderation** – the practice of posting UGC content immediately while putting copies of it in a queue for moderators to approve. This improves user satisfaction but increases the risk to institutions that offensive postings or misinformation will be seen
- **Community moderation** – This process depends on user communities to identify material that violates an institution's standards and is often used in conjunction with pre- and post-moderation
- **Distributed moderation** – This is a self-policing process that relies on the online community to determine what is acceptable content. Because this leaves institutions open to legal and reputational risks, some prefer an in-company distributed moderation system
- **Automated moderation** – the process of using digital tools such as AI that apply defined rules to reject or approve user submissions

All the types listed here, however, can be combined. In fact, a content moderation system can actually be less complex than many of these other tagging and flagging applications. Put simply, it consists of processes that set and enforce policies, along with a governance framework to manage potential bias and assure the safety of moderators. Operational excellence, including robust quality assurance, is also essential.

Here are the components of a successful content-moderation program:

Policy making

Delivering trustworthy and safe information is the overarching goal, but much depends on platform context and the market in which it operates. Each channel needs tailored policies that establish what's acceptable, so harmful content can be efficiently controlled, leaving legitimate information to flow freely. These policies must be clear and practical to enforce.

They must also be responsive to stakeholder expectations – and that calls for a feedback loop encompassing content creators, moderators, and others. The reason: constant feedback ensures that everyone involved understands the policies and has a voice in influencing them. Qualified teams need to evaluate this feedback, along with other information flowing from the system, to keep the policies and enforcement mechanisms aligned with changing user behavior.

Enforcement

AI plays an important role in detecting policy violations because it can quickly process vast amounts of information. AI can also help verify content producers, since trust and safety depends in part on who originated the information. But this is a highly sensitive and nuanced environment, so well-trained human recruits must still carry much of the burden. Their skill sets will vary, however, depending on the context. For example, the need for language proficiency and knowledge of cultural expression depends on the type of information involved.

Governance framework

With the scale of operations potentially running into thousands of employees, effective risk management requires solid governance. While individual initiative is expected, the system must function within a formal hierarchy of authorities, supported by close monitoring and reporting. This kind of governance ensures that the system meets its goals, which include guarding against unintentional bias that can lead to reputation disaster.

The governance framework should also protect the welfare of moderators. The daily job of reviewing and acting on unpleasant or offensive material can cause emotional harm, violating the employer's duty-of-care obligations and driving attrition. So controls must be put in place similar to those used for other health-and-safety hazards. This includes monitoring employee wellness and provision of counseling, as well as work rotation, systematic breaks as all tailored to specific operational, policy, and cultural environments.

Operational excellence

As with other mission-critical business systems, designing the right operating model for content moderation is essential. If the content is highly sensitive, companies might need 24x7 coverage so that enforcers can deal with potentially malicious material in near real time. In other settings, 24 hours may be an acceptable timeline for removing objectionable content. Either way, automation can optimize the process.

Operational excellence also demands a constantly evolving feedback loop. Is the enforcement team consistently making the right decisions? Should independent quality analysts review decisions? Are other changes needed, such as having more than one enforcer involved in every decision? Again, much depends on the nature of the information and cultural nuances.

The way forward

Genpact is working with leading content platforms to develop and refine state-of-the-art content moderation solutions. The systems we deliver now - AI, ML, sentiment analysis, tokenization analysis, and specialized searching - stringently meet today's needs. But they also continuously improve as the volume of content increases exponentially and new hazards appear.

Over time, technology will reduce the need for human intervention. Already, content-moderation systems are becoming better at predicting issues or viral events before they spread, facilitating rapid policy changes that reduce the risk of user exposure to harmful content.

To update their algorithms, AI and ML designers still need someone to identify content that gets through the system, including both false positives and false negatives. As such, humans will always be required to track and monitor moderated content.

There is a world of misinformation out there. No legitimate enterprise wants to contribute to it - however inadvertently. And with the right trust and safety partner, you can feel confident you won't.

This point of view was authored by Mark Hall, Sales Leader - High-Tech, Manufacturing and Services, Genpact.

About Genpact

Genpact (NYSE: G) is a global professional services firm that makes business transformation real. We drive digital-led innovation and digitally-enabled intelligent operations for our clients, guided by our experience running thousands of processes primarily for Global Fortune 500 companies. We think with design, dream in digital, and solve problems with data and analytics. Combining our expertise in end-to-end operations and our AI-based platform, Genpact Cora, we focus on the details - all 87,000+ of us. From New York to New Delhi and more than 25 countries in between, we connect every dot, reimagine every process, and reinvent companies' ways of working. We know that reimagining each step from start to finish creates better business outcomes. Whatever it is, we'll be there with you - accelerating digital transformation to create bold, lasting results - because transformation happens here.

For additional information visit, https://www.genpact.com/risk-compliance

Get to know us at Genpact.com and on LinkedIn, Twitter, YouTube, and Facebook.

genpact
Transformation Happens Here